(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0082948
(43) 공개일자 2020년07월08일

(51) 국제특허분류(Int. Cl.)
G06F 16/904 (2019.01) *G06F 16/906* (2019.01)
G06Q 40/06 (2012.01)
(52) CPC특허분류
G06F 16/904 (2019.01)
G06F 16/906 (2019.01)
(21) 출원번호 10-2018-0174052
(22) 출원일자 2018년12월31일
심사청구일자 2018년12월31일

(71) 출원인
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(72) 발명자
이주홍
경기도 부천시 신흥로 170-1 (중동, 위브더스테이트) 경기도 부천시 신흥로 170-1, 603동 803호 (중동,위브더스테이트)
안준규
충청북도 청주시 흥덕구 2순환로1340번길 72-11 (가경동) 가경푸르지오 503동 301호
(74) 대리인
이원희

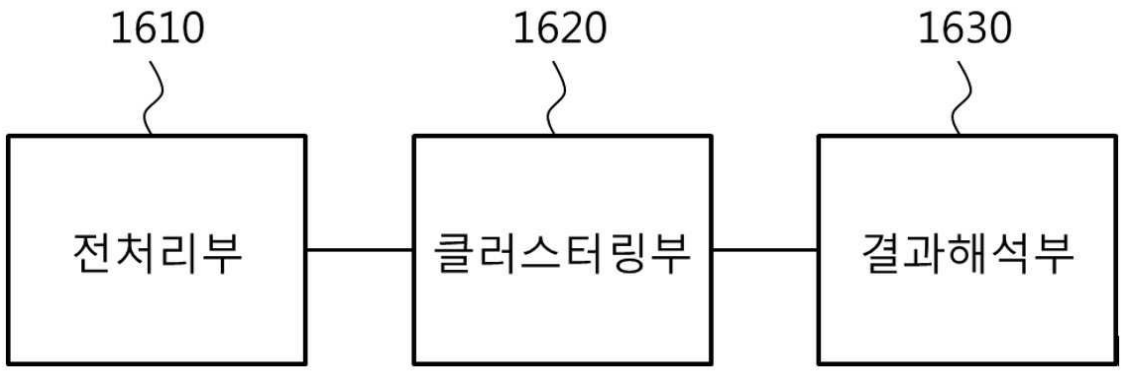
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 **공동성 관계가 있는 금융 시계열에 대한 클러스터링 방법 및 시스템**

(57) 요 약

본 발명은 공동성 관계가 있는 금융 시계열에 대한 클러스터링 방법 및 시스템에 있어서, 잡음을 제거하고 더 높은 공동성을 갖는 시계열 클러스터를 찾는 클러스터일 시스템 및 방법에 관한 것으로, 금융 데이터를 전처리 하는 단계; 전처리된 상기 데이터를 이용해 클러스터링 하는 단계; 및 상기 클러스트링 결과를 해석하는 단계;를 포함하는 구성을 개시한다.

대 표 도 - 도16



(52) CPC특허분류
G06Q 10/0637 (2013.01)
G06Q 40/06 (2013.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1345278188
부처명 교육부
연구관리전문기관 한국연구재단
연구사업명 기본연구(교육부)
연구과제명 [Ezbaro] 금융/경제 시계열 빅데이터 군집 알고리즘과 그 응용에 관한 연구
기 여 율 1/1
주관기관 인하대학교
연구기간 2018.03.01 ~ 2019.02.28

명 세 서

청구범위

청구항 1

금융 데이터를 전처리 하는 단계;

전처리된 상기 데이터를 이용해 클러스터링 하는 단계; 및

상기 클러스트링 결과를 해석하는 단계;를 포함하는 공동성 관계가 있는 금융 시계열에 대한 클러스터링 방법.

청구항 2

제1항에 있어서,

상기 전처리 단계는,

가중치를 적용하여 평균 스케일링하는 단계;를 포함하는 금융 시계열에 대한 클러스터링 방법.

청구항 3

제2항에 있어서,

상기 평균 스케일링 단계는,

하위 간격 계산 단계; 및

가중치 할당 단계;를 포함하는 금융 시계열에 대한 클러스터링 방법.

청구항 4

제3항에 있어서,

상기 전처리 단계는,

차원 축소 및 클러스터 직경 상한과 직경 한계를 추정하는 단계;를 더 포함하는 금융 시계열에 대한 클러스터링 방법.

청구항 5

제1항에 있어서,

상기 클러스터링 단계는,

사전 클러스터링 단계; 및

정제하여 클러스터링을 제어하는 단계;를 포함하는 금융 시계열에 대한 클러스터링 방법.

청구항 6

금융 데이터를 전처리 하는 전처리부;

전처리된 상기 데이터를 이용해 클러스터링 하는 클러스터링부; 및

상기 클러스트링 결과를 해석하는 결과해석부;를 포함하는 공동성 관계가 있는 금융 시계열에 대한 클러스터링 시스템.

청구항 7

제6항에 있어서,

상기 전처리부는,

금융 시계열 데이터를 가중치를 할당하여 평균 스케일링하는 포함하는 금융 시계열에 대한 클러스터링 시스템.

청구항 8

제7항에 있어서,

상기 전처리부는,

금융 시계열 데이터를 하위 간격 계산하고, 가중치 할당하는 가중 시계열 변환하는 금융 시계열에 대한 클러스터링 시스템.

청구항 9

제8항에 있어서,

상기 전처리부는,

차원 축소 및 클러스터 직경 상한과 직경 한계를 추정하는 금융 시계열에 대한 클러스터링 시스템.

청구항 10

제9항에 있어서,

상기 클러스터링부는,

사전 클러스터링을 수행하고, 정제하여 클러스터링을 제어하는 금융 시계열에 대한 클러스터링 시스템.

발명의 설명

기 술 분 야

[0001] 본 발명은 공동성 관계가 있는 금융 시계열에 대한 클러스터링 방법 및 시스템에 있어서, 잡음을 제거하고 더 높은 공동성을 갖는 시계열 클러스터를 찾는 클러스터일 시스템 및 방법에 관한 것이다.

배 경 기 술

[0003] 시계열은 시간에 따라 순차적으로 관찰되는 집합이다. 정보 기기의 발전으로 금융, 통신, 의약, 보건, 교통 등 다양한 분야에서 실시간으로 관측되는 시계열이 여러 분야의 응용 분야에 사용되고 있다. 그 중 금융 애플리케이션은 시계열이 가장 중요한 애플리케이션 중 하나다. 금융 응용 분야에는 환율, 통화량, 주가 지수, 주가, 상품 가격, 파생 상품 등이 있다. 따라서 금융 시계열 적용을 위한 효율적인 분석 방법을 개발하는 것에 있어서 금융 시계열의 중요한 기능을 고려하는 것이 매우 중요하다.

[0004] 금융 시계열은 랜덤 보행 특성을 가지고 있기 때문에 비정상적이다. 따라서 금융 시계열을 분석하여 투자 결정을 내리는 것은 매우 어렵다. 그러나 금융 시계열이 비정상적이라는 사실에도 불구하고, 두 개의 금융 시계열의 선형 결합의 결과인 Spread는 많은 상황에서 정상 상태를 만족시키는 것으로 밝혀졌다. 이 때, 두 개의 금융 시

계열은 공적분 관계라고 불리고, 같은 추세를 공유하는 것으로 간주된다. 공적분은 두 개의 비정상 시계열의 공동성 정도를 나타내는 척도로 사용된다. 공동성이 높은 금융 시계열 클러스터는 금융 시계열 응용 프로그램에 널리 적용될 수 있으므로 클러스터를 생성하는 클러스터링 알고리즘은 재무 응용 프로그램에서 중요한 연구 주제다. 이 결과는 금융 포트폴리오 선택 및 금융 정책 수립의 미래 가격을 예측하는 데 유용할 수 있다.

[0005] 기존의 시계열 클러스터링 알고리즘과 기존 금융 시계열 클러스터링 알고리즘의 주요 문제점은 다음과 같다. 첫째, 거리 계산 방법으로 Euclidean distance 와 Dynamic Time Warping 을 사용하지만, 거리 계산 방법은 시계열 데이터의 시간 가중치를 고려하지 않는다. 둘째, 금융 시계열의 추세 구성 요소 분석에서 과거 값에 비해 미래 가치의 증가 또는 감소 비율의 중요성을 무시한다. 셋째, 그들은 차원 감소 과정에서 너무 많은 정보 손실을 허용한다. 넷째, 그들은 낮은 정도의 공 동성 (co-movement)으로 너무 많은 소음을 포함하는 클러스터를 생성한다. 다섯째, 여러 클러스터에서 시계열을 중복 할당 할 수 없기 때문에 재무 분석에 유용한 정보가 손실 될 수 있다.

[0007] 예를 들어, 대한민국 공개특허 제102017-0078256호에는 시계열의 데이터를 예측하는 방법 및 그 장치가 개시되어 있고, 구체적으로는 트레이닝 기간 동안의 기 지정된 주기 단위의 측정치 시계열 데이터를 복수의 클러스터로 클러스터링 하는 단계; 상기 트레이닝 기간 동안의 복수의 환경 데이터를 수집하는 단계; 상기 복수의 환경 데이터 중 적어도 일부를 인자(factor)로 선택하는 단계; 상기 인자를 가리키는 축들로 구성되는 공간 또는 평면 상에서 상기 측정치 시계열 데이터의 클러스터를 최적으로 분류하는 분류 모델을 생성하는 단계; 상기 생성된 분류 모델의 성능 지표 값을 결정하는 단계; 상기 선택하는 단계, 상기 분류 모델을 생성하는 단계 및 상기 성능 지표 값을 결정하는 단계를, 상기 인자의 선택을 변경해 가면서 반복하여, 상기 성능 지표 값을 기준으로 상기 생성된 분류 모델 중 최적 분류 모델을 선정하는 단계; 및 상기 최적 분류 모델을 이용하여, 예측 기간의 상기 측정치 시계열 데이터의 클러스터를 예측 하는 단계를 포함하는, 시계열 데이터 예측 방법이 개시되어 있다. 그러나, 상기 클러스터링 방법은 공동성이 있는 시계열에 대한 클러스터링을 수행함에 있어 노이즈가 많이 발생할 수 있는 문제점이 있다.

[0009] 따라서, 상기한 문제를 해결할 수 있는 기술이 필요한 실정이다.

선행기술문헌

특허문헌

[0011] (특허문헌 0001) 대한민국 공개특허 제102017-0078256호

발명의 내용

해결하려는 과제

[0012] 따라서 본 발명은 상기의 문제점과 한계를 극복하기 위하여 기존의 클러스터링 방법과 비교할 때, 잡음을 제거하여 정확도를 높이고, 더 높은 공동성을 보이는 자료를 구별할 수 있는 클러스터링 방법 및 클러스터링 시스템을 제공하고자 한다.

과제의 해결 수단

[0014] 상기한 문제를 해결하기 위한 본 발명은 금융 데이터를 전처리 하는 단계;

[0015] 전처리된 상기 데이터를 이용해 클러스터링 하는 단계; 및 상기 클러스트링 결과를 해석하는 단계;를 포함하는 공동성 관계가 있는 금융 시계열에 대한 클러스터링 방법을 제공하고, 또한, 본 발명은 금융 데이터를 전처리하는 전처리부; 전처리된 상기 데이터를 이용해 클러스터링 하는 클러스터링부; 및 상기 클러스트링 결과를 해석하는 결과해석부;를 포함하는 공동성 관계가 있는 금융 시계열에 대한 클러스터링 시스템을 제공한다.

발명의 효과

[0017] 본 발명은 금융 시계열의 클러스터링 방법에 있어서 기존 방법과 대비할 때노이즈를 감소시키고 높은 공동성을 가지는 금융 시계열 클러스터링을 수행할 수 있다.

[0018] 본 발명은 기존의 시계열 클러스터링 알고리즘의 단점을 보완하고 높은 공 동성을 갖는 클러스터를 생성하는 클러스터링 방법 및 시스템을 제시한다.

[0019] 한편, 본 발명의 효과는 이상에서 언급한 효과들로 제한되지 않으며, 이하에서 설명할 내용으로부터 통상의 기술자에게 자명한 범위 내에서 다양한 효과들이 포함될 수 있다.

도면의 간단한 설명

[0021] 도 1은 밀도 기반 클러스터링의 문제점을 도시한 것이다.

도 2는 분할 기반 클러스터링의 문제점을 도시한 것이다.

도 3는 정규화 방법으로서의 Z- 변환의 문제를 도시한 그래프이다.

도 4는 대역 제한 신호, 하위 간격, 샘플링 된 포인트를 도시한 그래프이다.

도 5은 각 차원 축소 방법에 대한 DRE 결과 값이다.

도 6은 각 차원 축소 방법에 대한 DRaE 결과(α=50) 값이다.

도 7은 클러스터 지름의 변화를 도시한 그래프이다.

도 8은 직경 한계 추정 단계를 도시한 것이다.

도 9는 각 클러스터링 방법에 대한 RMSE의 결과이다.

도 10은 각 클러스터링 방법의 지름의 결과이다.

도 11은 각 클러스터링 방법에 대한 MSD의 결과이다.

도 12는 각 클러스터링 방법에 대한 ADF 테스트 결과이다.

도 13는 YADING으로 생성한 클러스터이다.

도 14는 3PTC로 생성한 클러스터이다.

도 15는 본 발명의 일 실시 예에 따른 방법으로 생성한 클러스터이다.

도 16은 본 발명의 일 실시 예에 따른 공동성 관계가 있는 금융 시계열에 대한 클러스터링 시스템의 블록도이다.

도 17은 본 발명의 일 실시 예에 따른 공동성 관계가 있는 금융 시계열에 대한 클러스터링 방법의 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0022] 이하, 첨부된 도면들을 참조하여 본 발명에 따른 '공동성 관계가 있는 금융 시계열에 대한 클러스터링 방법 및 시스템'을 상세하게 설명한다. 설명하는 실시 예들은 본 발명의 기술 사상을 당업자가 용이하게 이해할 수 있도록 제공되는 것으로 이에 의해 본 발명이 한정되지 않는다. 또한, 첨부된 도면에 표현된 사항들은 본 발명의 실시 예들을 쉽게 설명하기 위해 도식화된 도면으로 실제로 구현되는 형태와 상이할 수 있다.

[0023] 한편, 이하에서 표현되는 각 구성부는 본 발명을 구현하기 위한 예일 뿐이다. 따라서, 본 발명의 다른 구현에서는 본 발명의 사상 및 범위를 벗어나지 않는 범위에서 다른 구성부가 사용될 수 있다.

[0024] 또한, 각 구성부는 순전히 하드웨어 또는 소프트웨어의 구성만으로 구현될 수도 있지만, 동일 기능을 수행하는 다양한 하드웨어 및 소프트웨어 구성들의 조합으로 구현될 수도 있다. 또한, 하나의 하드웨어 또는 소프트웨어에 의해 둘 이상의 구성부들이 함께 구현될 수도 있다.

[0025] 또한, 어떤 구성요소들을 '포함'한다는 표현은, '개방형'의 표현으로서 해당 구성요소들이 존재하는 것을 단순히 지칭할 뿐이며, 추가적인 구성요소들을 배제하는 것으로 이해되어서는 안 된다.

[0026] 도 1은 밀도 기반 클러스터링의 문제점을 도시한 것이다.

[0027] 도 1을 참조하면, 시계열은 고차원적인 특성을 가지므로 시계열 분석을 위해 원시 시계열 데이터를 직접 사용하면 값 비싼 계산 및 차원의 폐해 문제가 발생할 수 있다. 시계열 분석에 사용되는 차원 축소 방법은 데이터 적응 방법과 비데이터 적응 방법으로 나눌 수 있다.

[0028] 데이터 적응 방식은 데이터의 크기가 가변적일 때 사용된다. PCA (Piecewise Constant Approximation), APCA (Adaptive Piecewise Constant Approximation), SAX (Symbolic Aggregate Approximation)를 포함할 수 있다. 비데이터 적응 방법은 데이터의 크기가 고정되어있을 때 사용될 수 있다. 여기에는 Random Mapping, PAA (Piecewise Aggregate Approximation), DFT (Discrete Fourier Transform), DCT (Discrete Cosine Transform)가 포함될 수 있다.

[0029] 기존 기술은 YADING이라 불리는 DBSCAN을 이용한 밀도 기반 클러스터링 알고리즘이 있다. 상기 기존 기술은 샘플링 방법을 사용하여 알고리즘의 실행 시간을 줄였다. 또한 그들은 변곡점을 사용하여 밀도 반경을 추정했다.

[0030] 다른 기존 기술로 시계열 클러스터링의 성능을 향상시키기 위해 다단계 클러스터링 알고리즘이 개발되었다. 2단계 클러스터링 방법으로 첫 번째 단계에서는 시계열 데이터가 SAX에 의해 변환된다. 그런 다음 CAST 알고리즘은 클러스터를 생성한다. 두 번째 단계에서는 각 클러스터의 시계열 데이터의 하위 시퀀스 정보를 사용하여 각 클러스터를 하위 클러스터로 나눈다.

[0031] 또 다른 기존 기술은 3PTC라고 불리는 3 단계 클러스터링 방법이 있다. 첫 번째 단계에서는 시계열 데이터가 SAX에 의해 변환된다. 그런 다음 K- 모드 알고리즘은 사전 클러스터를 생성한다. 두 번째 단계에서, 각 사전 클러스터는 PCS 알고리즘에 의해 서브 클러스터로 정제된다. 세 번째 단계에서는 형상 방법의 유사성을 이용하여 서브 클러스터 간의 유사도를 계산한다. 마지막으로 최종 클러스터는 하위 클러스터 병합으로 생성된다.

[0032] 금융 시계열 특성을 고려할 때 금융 시계열 클러스터링에서 중요한 요소로 고려해야 하는 핵심 요소는 클러스터 크기, 데이터 정규화, 거리 계산 및 시계열을 여러 클러스터에 중복 할당하는 것이다.

[0033] 기존의 시계열 클러스터링 방법은 클러스터의 적절한 크기에 대한 정보를 고려하지 않고 클러스터링을 수행한다. 클러스터의 적절한 크기가 고려되지 않은 경우 많은 양의 노이즈 데이터가 클러스터에 포함될 수 있다. 이러한 노이즈 데이터가 포함 된 클러스터를 사용하여 재무 데이터 분석을 수행하면 금융 데이터 분석의 결과가 신뢰할 수 없다.

[0034] 기존 기술인 밀도 기반 클러스터링은 데이터 공간에서 지속적으로 밀도가 높은 영역 내에서 임의의 모양으로 클러스터를 생성 할 수 있다. 해당 영역에서 생성된 클러스터의 크기는 어떤 모양에서도 매우 작거나 매우 클 수 있다. 클러스터 크기가 매우 큰 경우 클러스터 내의 가장 먼 두 데이터의 공동 이동 정도가 서로 매우 다를 수 있다. 도 1은 밀도 기반 클러스터의 문제점을 보여준다.

[0035] 도 2는 분할 기반 클러스터링의 문제점을 도시한 것이다.

[0036] 도 2를 참조하면, 분할 기반 클러스터링은 데이터 공간을 K 파티션으로 나누어 클러스터를 생성한다. 파티션의 수는 사용자가 결정할 수 있다. 사용자가 파티션 (K)의 수를 적절히 선택하지 않으면 가장 먼 두 데이터의 공동 이동 정도가 매우 다를 수 있다. 도 2는 파티션 기반 클러스터링의 문제점을 보여준다.

[0037] 도 3는 정규화 방법으로서의 Z- 변환의 문제를 도시한 그래프이다.

[0038] 도 3을 참조하면, 금융 시계열 추세 분석에서 과거 값에 비해 미래 가치가 얼마나 증가하는지 또는 감소 하는지를 알려주는 비율이 전체 시계열 데이터의 분포 및 분산보다 더 중요하다. 또한 금융 시계열은 각 데이터에 대해 매우 다른 규모를 가진다. 따라서 정규화 단계는 재무 시계열 분석에서 필수적이다. 그러나 대부분의 시계열 분석 방법은 시계열 데이터의 정규화를 고려하지 않거나 Z- 변환을 사용하여 정규화를 수행한다. Z- 변환은 다음 수학식 1과 같이 정의된다.

[0039] [수학식 1]

[0040] $$Y_t = \frac{X_t - \mu_t}{\sigma_X}$$

[0041] 여기서, Y_t는 시간 t에서 정규화 된 시계열을 나타내고, X_t는 시간 t에서 정규화되지 않은 시계열을 나타내며, μ_t는 X 데이터의 평균을 나타내며, σ_X는 X 데이터의 표준 편차를 나타낸다.

[0042] Z- 변환은 시계열 값의 표준 편차의 배수로 표현되기 때문에 시계열에 대한 정규화 방법으로 적합하지 않는다. 도 3은 각 시계열이 과거 값에 비해 미래 가치가 얼마나 증가하는지 또는 감소 하는지를 알려주는 비율이 다르더라도 두 시계열 데이터의 변경 비율이 유사 함을 보여준다.

[0043] 시계열 데이터는 장기간에 걸쳐 관찰된다. 시간이 지남에 따라 시계열을 생성하는 시스템의 상태가 변경 될 수 있으므로 시간 경과에 따라 시계열 분석 모델이 올바르게 작동하지 않을 수 있다. 따라서 모든 과거 데이터에 동일한 가중치를 부여하는 것보다 최근 데이터에 더 큰 가중치를 부여하는 것이 더 합리적이다. 특히, 이러한 현상은 금융 시계열에서 매우 빈번하게 발생하므로, 동심 유사성을 계산할 때 시간 가중치를 변경해야 한다.

[0044] 기존의 다단계 클러스터링 알고리즘은 사전 클러스터링 단계에서 사전 클러스터를 얻기 위해 축소된 차원의 시계열을 사용한다. 이 때 차원 축소 방법으로 SAX, PAA, DCT, DFT을 사용한다. 그러나 이러한 방법을 사용하면 정보가 과도하게 손실 될 수 있다.

[0045] 기존의 시계열 클러스터링 알고리즘은 여러 클러스터에 시계열을 중복 할당 할 수 없다. 그러나 하나의 시계열이 다른 클러스터에 포함된 시계열과 유사할 가능성이 있다. 따라서 시계열의 중복 할당을 허용하지 않으면 시계열 분석에서 중요한 정보가 손실 될 수 있다.

[0046] 이와 같은 문제점을 고려하여 본 발명에서는 금융 시계열 적용의 특성에 부합하는 효율적인 클러스터링 방법을 사용한다.

[0047] 도 4는 대역 제한 신호, 하위 간격, 샘플링 된 포인트를 도시한 그래프이다.

[0048] 도 4를 참조하면, 데이터 전처리에 있어서, 데이터 정규화를 위해 과거 가치 대비 미래 가치가 얼마나 증가하는지 또는 감소 하는지를 나타내는 비율을 나타내기 위해 평균 스케일링(Average Scaling)은 다음 수학식 2와 같이 정의된다.

[0049] [수학식 2]

[0050] $$Y_t = \frac{X_t}{\mu_X}$$

[0051] 여기서, Y_t는 시간 t에서의 정규화된 시계열을 나타내고, X_t는 시간 t에서의 시계열을 나타내고, μ_X는 X의 평균을 나타낸다.

[0052] 금융 시계열 분석에서 최근 데이터가 많을수록 향후 분석 결과에 영향을 미친다. 이 특성을 반영하기 위해 가중치 적용 시계열을 사용한다. 가중 시계열 변환은 가중치가 적용된 시계열을 생성한다. 이 변환은 다음의 두 단계를 포함할 수 있다: 하위 간격 계산 및 가중치 할당.

[0053] Nyquist Sampling Rate Condition과 Parseval Theorem을 기반으로 시간 구간의 하위 간격을 효율적으로 찾아내는 방법을 이용한다. Nyquist Sampling Rate 조건에서 샘플링 주파수가 최대 주파수의 2 배인 경우, 샘플링 포인트에 의한 주파수 앨리어싱(aliasing) 없이 대역 제한 신호를 완벽하게 재구성할 수 있다. 일부 시계열은 완벽한 대역 제한 신호가 아니지만 고주파 성분은 잡음으로 간주 될 수 있다. 따라서 대부분의 시계열은 대역 제한 신호로 간주 할 수 있다. Parseval 정리에 따르면, 시간 공간에서 시계열 값의 제곱의 합은 주파수 공간에서 주파수 성분의 계수의 제곱의 합과 같다. 따라서 Parseval 정리를 사용하여 원래 시계열의 에너지 대부분을 보존하는 최대 주파수 (Ω_max)를 찾는다(약 80-90 %). ($\Omega\_max = 2\pi k / N$, N은 시계열의 총 시간 간격에 해당하며, k는 주파수 1 / N의 다중 계수이다.)이 최대 주파수와 Nyquist Sampling Rate 조건을 사용하여 샘플링 기간 ($T_s \approx N/2k$, $2\pi/T_s \geq 2\Omega\_max$, $2\pi / T_s \geq 2 * 2\pi k / N$, $T_s \leq N / 2k$). 그런 다음 각 샘플링 된 기간을 시계열

의 하위 간격으로 해석한다. 도 4는 대역 제한 신호의 하위 간격을 보여준다.

[0054] Nyquist Sampling Rate Condition과 Parseval 정리에 의해 얻어진 부분 구간에 가중치를 선형적으로 할당한다. 예를 들어, 4 개의 서브 인터벌이 있는 경우, 과거로부터 최근까지의 시간 순서에 따라 1, 2, 3 및 4 가중치가 서브 인터벌에 할당된다.

[0055] 전처리 단계에서 차원 축소를 수행할 수 있다. DCT, PAA, PCA, SAX을 비교하여 차원 축소 방법을 선택 하였다. 차원 축소 방법으로 PCA (Principal Component Analysis)를 선택했다. PCA는 전체 데이터의 변동성을 가장 잘 나타내는 고유 벡터를 사용하여 데이터의 차원을 줄인다. PCA에 의해 감소된 차원의 수는 Scree 플롯을 사용하여 실험적으로 결정된다.

[0056] 전처리가 끝나면 클러스터링 단계를 수행할 수 있다.

[0057] 클러스터링 단계에서 사용되는 중요한 용어는 다음과 같이 정의된다.

[0058] 가중치 메트릭 공간 (WMS)

[0059] WMS는 다음과 같이 정의된다. X = {(f (t))|(f (t))는 (f (t)) 의 가중 된 시계열이고, t = 1, ..., T} SES (부분 구간 시계열의 유클리드 거리의 합) 거리.

[0060] (f (t))는 가중된 시계열 변환을 f (t)에 적용한 결과이다.

[0061] SES 거리

[0062] $$\sum_{i=1}^{m} SIDist_i$$

[0063] 여기서 $SIDist_i$는 부분 구간 시계열 f (t)의 i 번째 구간의 유클리드 거리이다.

[0064] RDS (Reduced Dimensional Space)

[0065] X = {g (z) | g (z)는 (f (t))의 의 축소 된 표현이다. T = 1, ..., T, z = 1, ..., rr << T}

[0066] g (z)는 (f (t))에 차원 감소 방법을 적용한 결과이다.

[0067] 가중치 공간에 사용된 클러스터의 지름 상한과 축소된 차원 공간에 사용된 클러스터의 지름 제한을 추정한다. 시계열은 추세 성분과 잡음 성분의 조합으로 간주 할 수 있다. 트렌드 구성 요소가 유사하고 노이즈가 충분히 적은 시계열을 찾아야 한다. 이를 위해 클러스터 직경을 제한한다. 사전 클러스터링 단계에서의 계산 효율성을 위해 금융 시계열을 축소된 차원 데이터로 변환한다. 그런 다음 축소된 차원 데이터에 대해 클러스터링이 수행된다. 그러나 치수 감소로 인해 사전 클러스터에 노이즈가 포함될 수 있다. 따라서 정제 단계에서 사전 클러스터를 정제 된 클러스터로 정제한다.

[0068] 직경 상한 추정

[0069] 정제 단계에서 클러스터의 크기를 제한하기 위해 직경 상한을 추정한다. 지경 상한은 클러스터 내의 시계열간의 SES_Distance의 최대 한도를 나타낸다. 지름 상한은 다음과 같이 추정된다. WMS에서 하나의 샘플을 추출한다. 그런 다음이 샘플과의 최단 거리 순서대로 이웃 집합에 시계열이 순차적으로 포함된다. 이웃 세트가 트렌드에서 크게 벗어나는 시계열을 가지고 있다면, 본 발명은 세트로부터 그 시계열을 제거하고, 최대 SES_Distance를 세트 내의 직경 후보로 고려한다. 여러 샘플을 반복적으로 작업하여 얻은 여러 직경 후보 값의 평균은 직경 상한으로 추정된다.

[0070] 직경 한계 추정

[0071] 사전 클러스터링 단계에서 클러스터의 크기를 제한하기 위해 사용되는 직경 한계는 다음과 같이 추정된다. S는 WMS에서 직경 상한(D)를 추정하는 데 사용되는 이웃 세트다. S의 직경 상한에 해당하는 두 개의 시계열이 있다. 본 발명은 축소된 차원 공간에서 유클리드 거리를 계산한다. 얻어진 거리를 직경 한계 후보(d ')로 한다. 축소된 차원 공간에는 정보가 손실된다. 클러스터의 최대 거리가 직경 제한 후보(d')로 제한되는 경우, 노이즈가 있는 데이터가 포함될 수 있을 뿐만 아니라 유사한 시계열이 동일한 클러스터에 포함되지 않을 수 있다. 그러므로 우리는 지름 제한값(d = γd', γ> 1)보다 약간 큰 값을 지름 한계 값으로 사용한다.

[0072] 사전 클러스터링 단계에서 ADupClustering 알고리즘은 Reduced Dimensional Space의 데이터에서 다음과 같이 수행된다. Agglomerative Clustering Algorithm은 클러스터를 생성한다. Agglomerative Clustering Algorithm에 의해 생성된 각 클러스터(R)에 대해 AdupClustering은 다른 클러스터에 포함된 시계열 중 클러스터 R에 포함되어야 하는 클러스터를 찾는다. 클러스터 R과의 거리는 지름 제한 (d)보다 작아야 한다. 즉, ADupClustering Algorithm은 하나의 시계열을 여러 클러스터에 할당한다. 즉, 중복 할당한다.

[0073] 사전 클러스터는 Reduced Dimensional Space에서 생성되기 때문에 서로 다른 추세를 갖는 시계열을 포함할 가능성이 있다. 정제 단계에서 각 사전 클러스터는 여러 개의 정제된 클러스터로 정제된다. 각각의 정제된 클러스터는 비슷한 경향 (높은 동조 정도)을 갖는 시계열만을 포함해야 한다. 이를 위해 ADupClustering 알고리즘은 지름 상한(D)을 사용하여 가중치 메트릭 공간에서 수행된다.

[0074] ADupClustering Algorithm에서 R을 집적 클러스터링에 의해 생성된 클러스터라고 가정한다. ρ를 R의 직경이라 하자. σ를 직경 상한(D) 또는 직경 한계(d)가 될 수 있는 직경 경계라고 하자. 그러면 ρ = dist (α, β)가되는 R에 α, β가 있다. ρ ≤ σ. c를 지름의 중심으로 놓는다((α β)). x = R이라하자. l = dist (x, c)라고하자. R '= R∪x 라하자. τ를 R '의 직경이라하자.

[0075] 보조 정리. l> (√3 σ) / 2이면 τ> σ

[0076] 증명. l> (√3 σ) / 2이고 $\overrightarrow{c\alpha}$ 와 $\overrightarrow{cx}$ 사이의 각도가 π / 2보다 크거나 같다면, dist (x, α)> $\sqrt{(\sigma/2)^2 + (\sqrt{3}\sigma/2)^2}$. $\tau > \sigma$, 왜냐하면 τ ≥dist (x, α)이기 때문이다. β의 경우 α의 경우와 동일하다.

[0077] 그러므로 l≤ (√3 σ) / 2 인 경우, τ는 지름 σ보다 작은 지 테스트해야 한다.

[0078] 본 발명의 성능을 증명하기 위해, 차원 축소의 효율성에 대한 실험, WMS (Weighted Metric Space)에서 Diameter Upperbound 결정에 대한 실험, RDS (Reduction Dimensional Space)에서 직경 한계 결정에 관한 실험, 우리의 방법을 다른 알고리즘과 비교하는 실험을 수행했다.

[0079] 한국 증시 (KOSPI, 코스닥) 및 UCR 시계열 아카이브의 주식 데이터를 실험 데이터로 사용 하였다. 한국의 주식 시장 데이터는 2016 년에 1,483 개의 주식으로 구성되며 각각의 주식은 4000 시간 규모를 가지고 있다. UCR 시계열 아카이브에서는 ShapsAll(512 차원), 단어 동의어(270 차원) 및 50 단어(270 차원)가 사용되었다. 실험은 Intel (R) Core (TM) i5-4570 CPU @ 3.20GHz, 4.00GB에서 수행되었다.

[0080] 본 발명에서 사용된 차원 축소 방법을 선택하기 위해 DCT, PAA, PCA를 비교한다. 차원 축소 효율성 (DRE) 및 차원 순위 효율성 (DRaE)이 비교 메트릭으로 사용된다. 차원 감소 효율(Dimension Reduction Efficiency)(DRE)는 다음 수학식 3과 같이 정의된다.

[0081] [수학식 3]

[0082] $$\frac{SSEW}{SSER} \times 100(\%)$$

[0083] SSEW (WMS의 제곱 오류 합계) 는 수학식 4와 같다.

[0084] [수학식 4]

[0085] $$\sum_{k \in SetW} \sum_{t=1}^{T} (f_k(t) - \mu(t))^2$$

[0086] 여기서 μ는 가중 된 메트릭 공간에서 임의로 선택된 기준 시계열을 나타내고, SetW는 기준 시계열에 가장 가까운 n 개의 시계열 집합이다.

[0087] 도 5은 각 치수 축소 방법에 대한 DRE 결과 값이다.

[0088] 도 5를 참조하면, SSER (RDS의 제곱 오류 합계)는 수학식 5와 같다.

[0089] [수학식 5]

[0090] $$\sum_{k \in SetR} \sum_{t=1}^{T} (f_k(t) - \mu(t))^2$$

[0091] 여기서 SetR은 n 개의 시계열 집합이다.

[0092] μ는 축소된 차원 공간에서 기준 시계열에 가장 근접한 시간 시계열이다. f_k, μ는 가중 된 메트릭 공간에서 표현된다.

[0093] DRE의 의미는 다음과 같다. SSER은 축소된 차원 공간에서 기준 시계열에 가까운 k 시계열을 사용하여 얻은 SSE (Square of Error) 값이다. 따라서 SSER는 항상 가중 측정 기준 공간에서 기준 시계열에 가까운 k 시계열을 사용하여 얻은 SSEW 값보다 크다. 결과적으로 정보 손실이 가장 적은 차원 축소 방법은 큰 DRE 값을 갖는다. 도 5는 PCA, DCT 및 PAA에 대한 DRE 값의 결과를 2 차원, 4 차원 및 6 차원으로 보여준다.

[0094] 도 6은 각 치수 축소 방법에 대한 DRaE 결과(α=50) 값이다.

[0095] 도 6을참조하면, 차원 순위 효율성 (DRaE)은 다음 수학식 6과 같이 정의된다.

[0096] [수학식 6]

[0097] $$\frac{\alpha N}{\sum_{k=1}^{N} |r(d_k) - rr(d_k)| + \alpha N} \times 100(\%)$$

[0098] 여기서, N은 SetW에 포함 된 시계열의 수를 나타낸다. d_k는 가중된 메트릭 공간에서 μ에서 k 번째 이웃을 나타낸다. $r(d_k)$는 가중 메트릭 공간의 기준 시계열에서 d_k의 순위다. $r(d_k) = k$. $rr(d_k)$는 축소된 차원 공간에서 기준 시계열에서 d_k의 순위이다.

[0099] DRaE의 의미는 다음과 같다. $r(d_k) = rr(d_k)$이면 차원 축소 방법으로 인해 순위 정보가 손실되지 않으므로 순위 오류가 발생하지 않는다. 따라서 DRaE의 가치는 100 %이다. 그러나 $r(d_k) \neq rr(d_k)$인 경우 차원 축소 방법으로 인해 순위 정보가 손실되므로 순위 오류가 발생한다. 따라서 DRaE의 가치는 100 %보다 낮다. 도 6은 PCA, DCT 및 PAA에 대한 DRaE 결과를 2,4,6 차원에서 보여준다.

[0100] 도 7은 클러스터 지름의 변화를 도시한 그래프이다.

[0101] 도 7을 참조하면, 본 발명은 PCA의 DRE와 DRaE 값이 DCT와 PAA의 그것보다 더 큰 것을 확인했다. 결과적으로 본 발명에서는 차원 축소 방법으로 PCA를 선택했다.

[0102] 도 7은 직경 상한 추정에 대한 실험을 보여준다. 본 발명은 N 개의 이웃을 기준 시계열 μ로부터 이웃 집합에 순차적으로 추가하고 이웃 집합 내의 데이터 사이의 최대 거리 (Diameter)의 변화를 식별한다. 특정 최대 거리에서 이웃 집합의 대부분의 시계열과 매우 다른 추세를 갖는 시계열이 나타난다. 최대 거리는 직경 상한의 후보다.

[0103] 도 8은 직경 한계 추정 단계를 도시한 것이다.

[0104] 도 8을 참조하면, 감소된 차원 공간에서 사용 된 직경 한계를 추정하는 프로세스를 도시한다. 오른쪽 그래프는 가중 메트릭 공간에서 기준 시계열의 이웃 집합이 순차적으로 이웃 집합에 추가 될 때 이웃 집합의 지름이 증가한다는 것을 보여준다. 왼쪽 그래프는 다음과 같이 나타낸다. 축소된 차원 공간에서 기준 시계열의 이웃이 순차적으로 이웃 세트에 추가되면 이웃 세트의 지름이 증가한다. 직경 제한은 다음과 같이 추정된다. 오른쪽 그래프에서 특정 직경 상한을 찾은 다음 특정 직경 상한에 해당하는 k 번째 시간 시리즈를 찾는다. 다음으로, 왼쪽 그래프에서 k 번째 시간 시리즈에 해당하는 지름을 찾는다. 본 발명은 왼쪽 그래프에서 구한 k 번째 시계열에 해당하는 지름을 지름 제한 후보로 간주한다. 직경 제한 후보로 직경 한계 후보 × γ를 사용한다.

[0105] 도 9는 각 클러스터링 방법에 대한 RMSE의 결과이고, 도 10은 각 클러스터링 방법의 지름의 결과이고, 도 11은 각 클러스터링 방법에 대한 MSD의 결과이다.

[0106] 도 9 내지 도 11을 참조하면, 본 발명의 일 실시 예에 따른 방법을 다른 알고리즘과 비교하기 위해 가장 최근에

발표된 기술인 3PTC와 YADING을 선택했다. 3PTC는 금융 시계열 클러스터링에 대한 최신 기술이다. YADING은 밀도 기반 시계열 클러스터링에 관한 최신 기술이다. 본 발명은 두 개의 평가 방법을 사용하여 클러스터에서 시계열의 공동 이동 정도를 평가한다.

[0107] RMSE(root mean square error)는 다음 수학식 7과 같이 정의된다.

[0108] [수학식 7]

[0109] $$\sqrt{\frac{1}{|cluster|}\sum_{i\in cluster}\sum_{t=1}^{n}(f_i(t)-p(t))^2}$$

[0110] 여기서, |cluster|는 클러스터의 시계열 수를 나타낸다. $f_i(t)$는 클러스터의 i 번째 시계열을 나타낸다. P(t)는 클러스터의 프로토타입 시계열을 나타낸다.

[0111] 프로토 타입 시계열 p (t)는 다음 수학식 8과 같이 정의된다.

[0112] [수학식 8]

[0113] $$\frac{1}{|cluster|}\sum_{i\in cluster} f_i(t)$$

[0114] MSD(최대 표준 편차)는 다음 수학식 9와 같이 정의된다.

[0115] [수학식 9]

[0116] $$\max_{i\in cluster}\sqrt{\frac{1}{n}\sum_{t=1}^{n}(f_i(t)-p(t))^2}$$

[0117] RMSE는 클러스터의 일관성을 평가하는 척도다. 따라서 값이 작을수록 클러스터의 공동 이동이 높다. 그러나 RMSE는 평균값이기 때문에 상당히 다른 동조 정도를 갖는 시계열이 클러스터에 포함되는 경우를 구별 할 수 없다. MSD 및 직경은 최대 값의 지표다. 클러스터 내의 일관성 정도를 측정하지는 않지만 클러스터 내에서 시계열의 최대 변동을 측정 할 수 있다. 따라서 클러스터의 동공 이동 정도를 평가하기 위해서는 동시에 세 값을 작게 해야 한다.

[0118] 도 9 내지 도 11에 도시 된 바와 같이, 본 방법의 3 가지 값은 다른 방법보다 작다. 따라서 본 발명의 방법으로 생성된 클러스터의 공동 운동(co-movement) 정도가 가장 높음을 알 수 있다.

[0119] 도 12는 각 클러스터링 방법에 대한 ADF 테스트 결과이다.

[0120] 도 12를 참조하면, 공적분은 공운동의 지표로 사용될 수 있다. 따라서 본 발명은 공적분 평가 방법 (ADF test)을 사용하여 클러스터에서 시계열의 공 동성 정도를 평가한다.

[0121] 알고리즘에 의해 생성된 클러스터의 시계열간에 공적분 관계가 있는지 확인하기 위해 클러스터의 모든 시계열 쌍에 대해 ADF 테스트를 수행한다. 그런 다음 P 값을 얻는다. P 값이 임계 값보다 낮으면 귀무 가설(null hypothesis)은 기각된다. 이것은 두 개의 시계열이 공적분 관계에 있을 확률이 높다는 것을 의미한다. ADF 테스트로 얻은 P 값으로부터 귀무 가설을 기각한 비율을 구한다. 비율이 높을수록 클러스터의 시계열 쌍이 공적분 관계를 만족시킨다. 따라서 우리는 클러스터에서 시계열의 동일 이동 정도를 볼 수 있다. 도 12의 결과는 다음을 의미한다.

[0122] 본 발명의 일 실시 예에 따른 방법으로 생성된 클러스터에 대한 귀무 가설의 거부율이 YADING, 3PTC의 거부율보다 크다는 것을 발견했다. 이것은 본 발명의 방법에 의해 생성된 클러스터의 시계열 쌍이 다른 방법에 의해 생성된 클러스터의 시계열 쌍보다 높은 동조 정도에 있다는 것을 의미한다.

[0123] 도 13는 YADING으로 생성한 클러스터이고, 도 14는 3PTC로 생성한 클러스터이고, 도 15는 본 발명의 일 실시 예에 따른 방법으로 생성한 클러스터이다.

[0124] 도 13 내지 15는 클러스터링 알고리즘 평가에 사용 된 각 알고리즘의 대표 클러스터를 개시한다.

[0125] 도 13, 도 14는 클러스터 내에 다양한 경향이 있는 하위 클러스터가 있음을 알 수 있다. YADING, 3PTC에 의해 생성된 클러스터의 시계열 쌍은 고도의 공 동성을 만족시키지 못한다.

[0126] 도 15는 본 발명의 방법의 결과를 도시한다. 우리의 방법으로 생성된 클러스터의 시계열 쌍이 높은 수준의 공동성을 만족한다는 것을 알 수 있다.

[0127] 본 발명에서는 공 동성이 높은 금융 시계열 클러스터를 찾는 방법을 개시한다. 특히, 본 발명은 금융 시계열의 특성을 고려한다. 본 발명에서 제안한 클러스터링 알고리즘은 포트폴리오 선택, 위험 관리, 자산 할당, 시계열 예측과 같은 다양한 금융 투자에 적용될 수 있다. 또한 거시 경제적 용도로 사용될 가능성이 높기 때문에 우리의 방법이 매우 가치 있다고 확신한다.

[0128] 도 16은 본 발명의 일 실시 예에 따른 공동성 관계가 있는 금융 시계열에 대한 클러스터링 시스템의 블록도이다.

[0129] 도 16을 참조하면, 본 발명의 일 실시 예에 따른 공동성 관계가 있는 금융 시계열에 대한 클러스터링 시스템은 전처리부(1610), 클러스터링부(1620) 및 결과해석부(1630)를 포함할 수 있다.

[0130] 상기 전처리부(1610)는 데이터 정규화를 위해 과거 가치 대비 미래 가치가 얼마나 증가하는지 또는 감소 하는지를 나타내는 비율을 나타내기 위해 평균 스케일링(Average Scaling)을 수행할 수 있다. 상기 평균 스케일링은 다음 수학식 2와 같이 정의된다.

[0131] [수학식 2]

[0132] $$Y_t = \frac{X_t}{\mu_X}$$

[0133] 여기서, Y_t는 시간 t에서의 정규화된 시계열을 나타내고, X_t는 시간 t에서의 시계열을 나타내고, μ_X는 X의 평균을 나타낸다.

[0134] 상기 전처리부(1610)에서 금융 시계열 분석에서 최근 데이터가 많을수록 향후 분석 결과에 영향을 미친다. 이 특성을 반영하기 위해 가중치 적용 시계열을 사용한다. 가중 시계열 변환은 가중치가 적용된 시계열을 생성한다. 이 변환은 다음의 두 단계를 포함할 수 있다: 하위 간격 계산 및 가중치 할당.

[0135] 상기 전처리부(1610)는 Nyquist Sampling Rate 조건과 Parseval 정리를 기반으로 시간 구간의 하위 간격을 효율적으로 찾아내는 방법을 이용한다. Nyquist Sampling Rate 조건에서 샘플링 주파수가 최대 주파수의 2 배인 경우, 샘플링 포인트에 의한 주파수 앨리어싱(aliasing) 없이 대역 제한 신호를 완벽하게 재구성할 수 있다. 일부 시계열은 완벽한 대역 제한 신호가 아니지만 고주파 성분은 잡음으로 간주 될 수 있다. 따라서 대부분의 시계열은 대역 제한 신호로 간주 할 수 있다. Parseval 정리에 따르면, 시간 공간에서 시계열 값의 제곱의 합은 주파수 공간에서 주파수 성분의 계수의 제곱의 합과 같다. 따라서 Parseval 정리를 사용하여 원래 시계열의 에너지 대부분을 보존하는 최대 주파수 (Ω_max)를 찾는다(약 80-90 %). ($\Omega\_max = 2\pi k / N$, N은 시계열의 총 시간 간격에 해당하며, k는 주파수 1 / N의 다중 계수이다.)이 최대 주파수와 Nyquist Sampling Rate 조건을 사용하여 샘플링 기간($T_s \approx N/2k$, $2\pi/T_s \geq 2\Omega\_max$, $2\pi / T_s \geq 2 * 2\pi k / N$, $T_s \leq N / 2k$). 그런 다음 각 샘플링 된 기간을 시계열의 하위 간격으로 해석한다. 도 4는 대역 제한 신호의 하위 간격을 보여준다.

[0136] 상기 전처리부(1610)는 Nyquist Sampling Rate 조건과 Parseval 정리에 의해 얻어진 부분 구간에 가중치를 선형적으로 할당한다. 예를 들어, 4 개의 서브 인터벌이 있는 경우, 과거로부터 최근까지의 시간 순서에 따라 1, 2, 3 및 4 가중치가 서브 인터벌에 할당된다.

[0137] 상기 전처리부(1610)는 전처리 단계에서 차원 축소를 수행할 수 있다. DCT, PAA, PCA, SAX을 비교하여 차원 축소 방법을 선택 하였다. 차원 축소 방법으로 PCA (Principal Component Analysis)를 선택했다. PCA는 전체 데이터의 변동성을 가장 잘 나타내는 고유 벡터를 사용하여 데이터의 차원을 줄인다. PCA에 의해 감소된 치수의 수는 Scree 플롯을 사용하여 실험적으로 결정된다.

[0138] 상기 클러스터링부(1620)는 가중치 공간에 사용된 클러스터의 지름 상한과 축소된 차원 공간에 사용된 클러스터의 지름 제한을 추정한다. 시계열은 추세 성분과 잡음 성분의 조합으로 간주할 수 있다. 트렌드 구성 요소가 유

사하고 노이즈가 충분히 적은 시계열을 찾아야 한다. 이를 위해 클러스터 직경을 제한한다. 사전 클러스터링 단계에서의 계산 효율성을 위해 금융 시간 시리즈를 축소된 차원 데이터로 변환한다. 그런 다음 축소된 차원 데이터에 대해 클러스터링이 수행된다. 그러나 차원 감소로 인해 사전 클러스터에 노이즈가 포함될 수 있다. 따라서 정제 단계에서 사전 클러스터를 정제된 클러스터로 정제한다.

[0139] 상기 클러스터링부(1620)는 직경 상한 추정을 수행할 수 있다.

[0140] 상기 클러스터링부(1620)는 정제 단계에서 클러스터의 크기를 제한하기 위해 직경 상한을 추정한다. 지경 상한은 클러스터 내의 시계열간의 SES_Distance의 최대 한도를 나타낸다. 지름 상한은 다음과 같이 추정된다. WMS에서 하나의 샘플을 추출한다. 그런 다음이 샘플과의 최단 거리 순서대로 이웃 집합에 시계열이 순차적으로 포함된다. 이웃 세트가 트렌드에서 크게 벗어나는 시계열을 가지고 있다면, 본 발명은 세트로부터 그 시계열을 제거하고, 최대 SES_Distance를 세트 내의 직경 후보로 고려한다. 여러 샘플을 반복적으로 작업하여 얻은 여러 직경 후보 값의 평균은 직경 상한으로 추정된다.

[0141] 상기 클러스터링부(1620)는 직경 한계 추정할 수 있다.

[0142] 상기 클러스터링부(1620)는 사전 클러스터링 단계에서 클러스터의 크기를 제한하기 위해 사용되는 직경 한계는 다음과 같이 추정된다. S는 WMS에서 직경 상한(D)를 추정하는 데 사용되는 이웃 세트다. S의 직경 상한에 해당하는 두 개의 시계열이 있다. 본 발명은 축소된 차원 공간에서 유클리드 거리를 계산한다. 얻어진 거리를 직경 한계 후보(d')로 한다. 축소된 차원 공간에는 정보가 손실된다. 클러스터의 최대 거리가 직경 제한 후보(d')로 제한되는 경우, 노이즈가 있는 데이터가 포함될 수 있을 뿐만 아니라 유사한 시계열이 동일한 클러스터에 포함되지 않을 수 있다. 그러므로 우리는 지름 제한값(d = ɣd', ɣ> 1)보다 약간 큰 값을 지름 한계 값으로 사용한다.

[0143] 상기 클러스터링부(1620)는 사전 클러스터링 단계에서 ADupClustering 알고리즘은 Reduced Dimensional Space의 데이터에서 다음과 같이 수행된다. Agglomerative Clustering Algorithm은 클러스터를 생성한다. Agglomerative Clustering Algorithm에 의해 생성된 각 클러스터(R)에 대해 AdupClustering은 다른 클러스터에 포함된 시계열 중 클러스터 R에 포함되어야 하는 클러스터를 찾는다. 클러스터 R과의 거리는 지름 제한 (d)보다 작아야 한다. 즉, ADupClustering Algorithm은 하나의 시계열을 여러 클러스터에 할당한다. 즉, 중복 할당한다.

[0144] 상기 클러스터링부(1620)는 사전 클러스터는 Reduced Dimensional Space에서 생성되기 때문에 서로 다른 추세를 갖는 시계열을 포함할 가능성이 있다. 정제 단계에서 각 사전 클러스터는 여러 개의 정제된 클러스터로 정제된다. 각각의 정제된 클러스터는 비슷한 경향 (높은 동조 정도)을 갖는 시계열만을 포함해야 한다. 이를 위해 ADupClustering 알고리즘은 지름 상한(D)을 사용하여 가중치 메트릭 공간에서 수행된다.

[0145] 상기 결과해석부(1630)는 상기 클러스터링 결과를 이용해 금융 시계열, 포트폴리오 선택 및 금융 정책 수립의 미래 가격을 예측할 수 있다.

[0146] 도 17은 본 발명의 일 실시 예에 따른 공동성 관계가 있는 금융 시계열에 대한 클러스터링 방법의 흐름도이다.

[0147] 도 17을 참조하면, 본 발명의 일 실시 예에 따른 공동성 관계가 있는 금융 시계열에 대한 클러스터링 방법은 금융 데이터를 전처리 하는 단계(S1710)를 포함할 수 있다.

[0148] S1710 단계에서, 상기 전처리부(1610)는 데이터 정규화를 위해 과거 가치 대비 미래 가치가 얼마나 증가하는지 또는 감소 하는지를 나타내는 비율을 나타내기 위해 평균 스케일링(Average Scaling)을 수행할 수 있다. 상기 평균 스케일링은 다음 수학식 2와 같이 정의된다.

[0149] [수학식 2]

[0150] $$Y_t = \frac{X_t}{\mu_X}$$

[0151] 여기서, Y_t는 시간 t에서의 정규화된 시계열을 나타내고, X_t는 시간 t에서의 시계열을 나타내고, μ_X는 X의 평균을 나타낸다.

[0152] S1710 단계에서, 상기 전처리부(1610)에서 금융 시계열 분석에서 최근 데이터가 많을수록 향후 분석 결과에 영향을 미친다. 이 특성을 반영하기 위해 가중치 적용 시계열을 사용한다. 가중 시계열 변환은 가중치가 적용된

시계열을 생성한다. 이 변환은 다음의 두 단계를 포함할 수 있다: 하위 간격 계산 및 가중치 할당.

[0153] S1710 단계에서, 상기 전처리부(1610)는 Nyquist Sampling Rate 조건과 Parseval 정리를 기반으로 시간 구간의 하위 간격을 효율적으로 찾아내는 방법을 이용한다. Nyquist Sampling Rate 조건에서 샘플링 주파수가 최대 주파수의 2 배인 경우, 샘플링 포인트에 의한 주파수 앨리어싱(aliasing) 없이 대역 제한 신호를 완벽하게 재구성할 수 있다. 일부 시계열은 완벽한 대역 제한 신호가 아니지만 고주파 성분은 잡음으로 간주 될 수 있다. 따라서 대부분의 시계열은 대역 제한 신호로 간주 할 수 있다. Parseval 정리에 따르면, 시간 공간에서 시계열 값의 제곱의 합은 주파수 공간에서 주파수 성분의 계수의 제곱의 합과 같다. 따라서 Parseval 정리를 사용하여 원래 시계열의 에너지 대부분을 보존하는 최대 주파수 (Ω_max)를 찾는다(약 80-90 %). ($\Omega\_max = 2\pi k / N$, N은 시계열의 총 시간 간격에 해당하며, k는 주파수 1 / N의 다중 계수이다.)이 최대 주파수와 Nyquist Sampling Rate 조건을 사용하여 샘플링 기간($T_s \approx N/2k$, $2\pi/T_s \geq 2\Omega\_max$, $2\pi / T_s \geq 2 * 2\pi k / N$, $T_s \leq N / 2k$). 그런 다음 각 샘플링 된 기간을 시계열의 하위 간격으로 해석한다. 도 4는 대역 제한 신호의 하위 간격을 보여준다.

[0154] S1710 단계에서, 상기 전처리부(1610)는 Nyquist Sampling Rate 조건과 Parseval 정리에 의해 얻어진 부분 구간에 가중치를 선형적으로 할당한다. 예를 들어, 4 개의 서브 인터벌이 있는 경우, 과거로부터 최근까지의 시간 순서에 따라 1, 2, 3 및 4 가중치가 서브 인터벌에 할당된다.

[0155] S1710 단계에서, 상기 전처리부(1610)는 전처리 단계에서 차원 축소를 수행할 수 있다. DCT, PAA, PCA, SAX을 비교하여 차원 축소 방법을 선택 하였다. 차원 축소 방법으로 PCA (Principal Component Analysis)를 선택했다. PCA는 전체 데이터의 변동성을 가장 잘 나타내는 고유 벡터를 사용하여 데이터의 차원을 줄인다. PCA에 의해 감소된 치수의 수는 Scree 플롯을 사용하여 실험적으로 결정된다.

[0156] 본 발명의 일 실시 예에 따른 공동성 관계가 있는 금융 시계열에 대한 클러스터링 방법은 전처리된 상기 데이터를 이용해 클러스터링 하는 단계(S1720)를 포함할 수 있다.

[0157] S1720 단계에서, 상기 클러스터링부(1620)는 가중치 공간에 사용된 클러스터의 지름 상한과 축소된 차원 공간에 사용된 클러스터의 지름 제한을 추정한다. 시계열은 추세 성분과 잡음 성분의 조합으로 간주할 수 있다. 트렌드 구성 요소가 유사하고 노이즈가 충분히 적은 시계열을 찾아야 한다. 이를 위해 클러스터 직경을 제한한다. 사전 클러스터링 단계에서의 계산 효율성을 위해 금융 시간 시리즈를 축소된 차원 데이터로 변환한다. 그런 다음 축소된 차원 데이터에 대해 클러스터링이 수행된다. 그러나 치수 감소로 인해 사전 클러스터에 노이즈가 포함될 수 있다. 따라서 정제 단계에서 사전 클러스터를 정제 된 클러스터로 정제한다.

[0158] S1720 단계에서, 상기 클러스터링부(1620)는 직경 상한 추정을 수행할 수 있다.

[0159] 상기 클러스터링부(1620)는 정제 단계에서 클러스터의 크기를 제한하기 위해 직경 상한을 추정한다. 직경 상한은 클러스터 내의 시계열간의 SES_Distance의 최대 한도를 나타낸다. 지름 상한은 다음과 같이 추정된다. WMS에서 하나의 샘플을 추출한다. 그런 다음이 샘플과의 최단 거리 순서대로 이웃 집합에 시계열이 순차적으로 포함된다. 이웃 세트가 트렌드에서 크게 벗어나는 시계열을 가지고 있다면, 본 발명은 세트로부터 그 시계열을 제거하고, 최대 SES_Distance를 세트 내의 직경 후보로 고려한다. 여러 샘플을 반복적으로 작업하여 얻은 여러 직경 후보 값의 평균은 직경 상한으로 추정된다.

[0160] S1720 단계에서, 상기 클러스터링부(1620)는 직경 한계 추정할 수 있다.

[0161] 상기 클러스터링부(1620)는 사전 클러스터링 단계에서 클러스터의 크기를 제한하기 위해 사용되는 직경 한계는 다음과 같이 추정된다. S는 WMS에서 직경 상한(D)를 추정하는 데 사용되는 이웃 세트다. S의 직경 상한에 해당하는 두 개의 시계열이 있다. 본 발명은 축소된 차원 공간에서 유클리드 거리를 계산한다. 얻어진 거리를 직경 한계 후보(d')로 한다. 축소된 차원 공간에는 정보가 손실된다. 클러스터의 최대 거리가 직경 제한 후보(d')로 제한되는 경우, 노이즈가 있는 데이터가 포함될 수 있을 뿐만 아니라 유사한 시계열이 동일한 클러스터에 포함되지 않을 수 있다. 그러므로 우리는 지름 제한값($d = \gamma d'$, $\gamma > 1$)보다 약간 큰 값을 지름 한계 값으로 사용한다.

[0162] S1720 단계에서, 상기 클러스터링부(1620)는 사전 클러스터링 단계에서 ADupClustering 알고리즘은 Reduced Dimensional Space의 시계열에서 다음과 같이 수행된다. Agglomerative Clustering Algorithm은 클러스터를 생성한다. Agglomerative Clustering Algorithm에 의해 생성된 각 클러스터(R)에 대해 AdupClustering은 다른 클러스터에 포함된 시계열 중 클러스터 R에 포함되어야 하는 클러스터를 찾는다. 클러스터 R과의 거리는 지름 제한 (d)보다 작아야 한다. 즉, ADupClustering Algorithm은 하나의 시계열을 여러 클러스터에 할당한다. 즉, 중

복 할당한다.

[0163] S1720 단계에서, 상기 클러스터링부(1620)는 사전 클러스터는 Reduced Dimensional Space에서 생성되기 때문에 서로 다른 추세를 갖는 시계열을 포함할 가능성이 있다. 정제 단계에서 각 사전 클러스터는 여러 개의 정제된 클러스터로 정제된다. 각각의 정제된 클러스터는 비슷한 경향 (높은 동조 정도)을 갖는 시계열만을 포함해야 한다. 이를 위해 ADupClustering 알고리즘은 지름 상한(D)을 사용하여 가중치 메트릭 공간에서 수행된다.

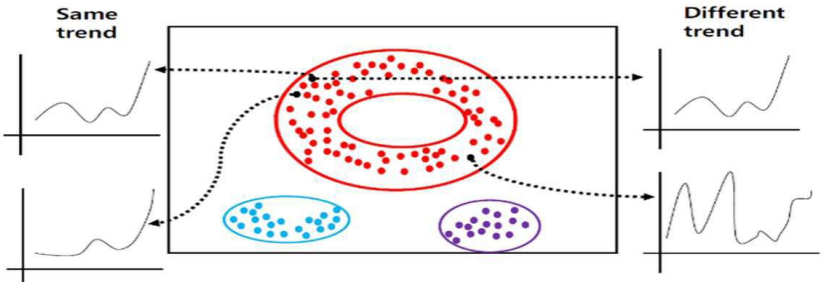
[0164] 본 발명의 일 실시 예에 따른 공동성 관계가 있는 금융 시계열에 대한 클러스터링 방법은 상기 클러스트링 결과를 해석하는 단계(S1730)를 포함할 수 있다.

[0165] S1730 단계에서, 상기 결과해석부(1630)는 상기 클러스터링 결과를 이용해 금융 시계열, 포트폴리오 선택 및 금융 정책 수립의 미래 가격을 예측할 수 있다.

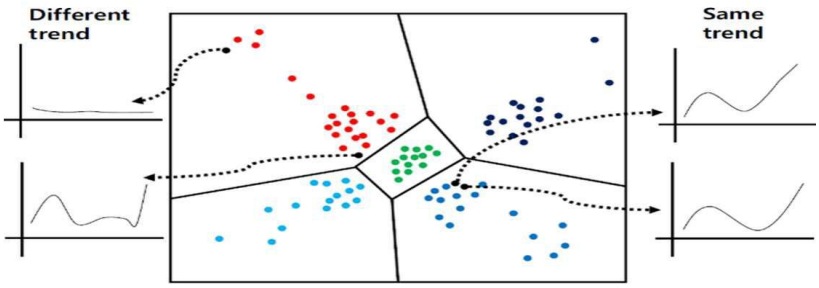
[0167] 이제까지 본 발명에 대하여 그 바람직한 실시 예들을 중심으로 살펴보았다. 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자는 본 발명이 본 발명의 본질적인 특성에서 벗어나지 않는 범위에서 변형된 형태로 구현될 수 있음을 이해할 수 있을 것이다. 그러므로 개시된 실시 예들은 한정적인 관점이 아니라 설명적인 관점에서 고려되어야 한다. 본 발명의 범위는 전술한 설명이 아니라 특허청구범위에 나타나 있으며, 그와 동등한 범위 내에 있는 모든 차이점은 본 발명에 포함된 것으로 해석되어야 할 것이다.

도면

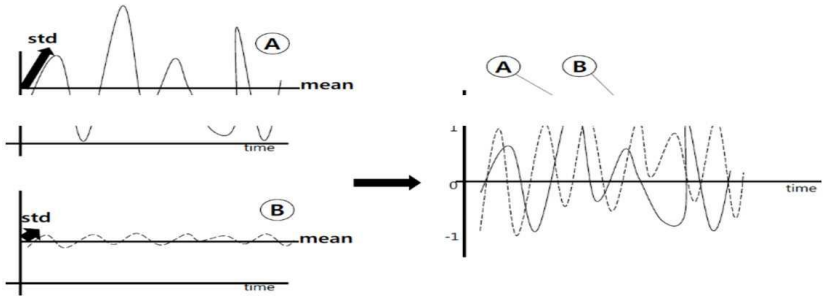
도면1



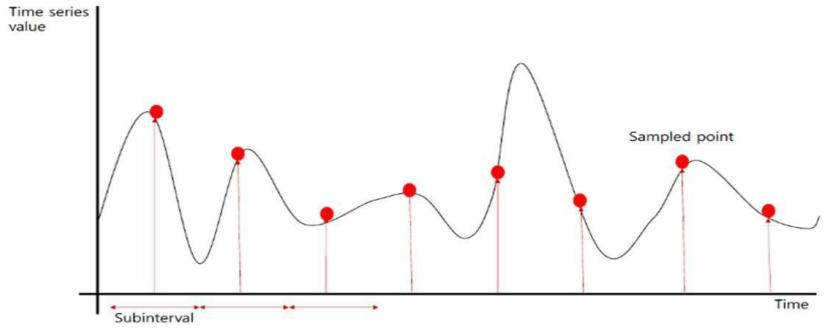
도면2



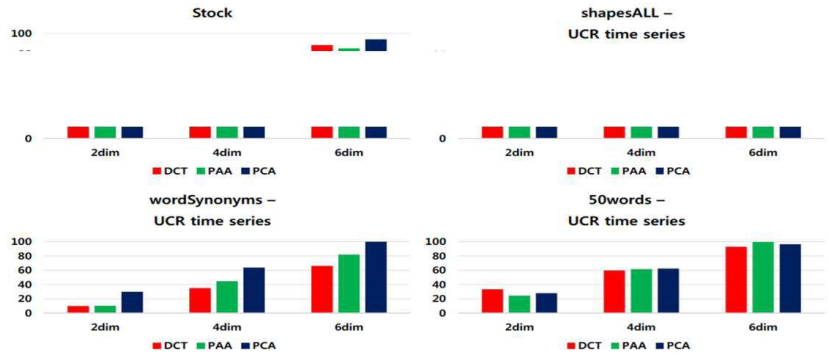
도면3



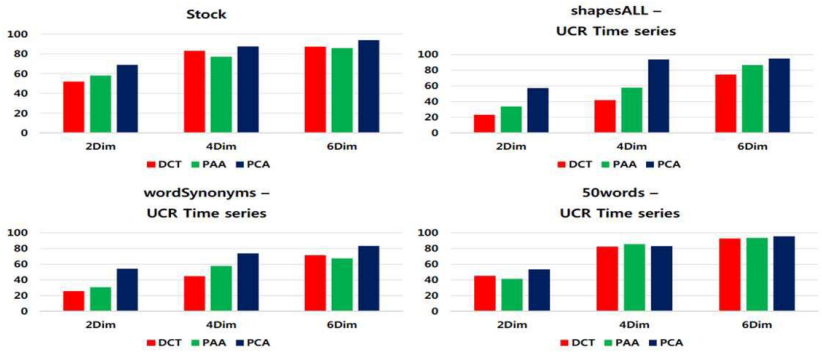
도면4



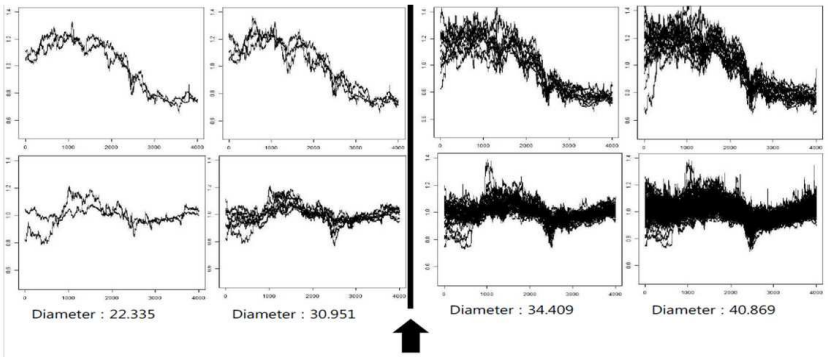
도면5



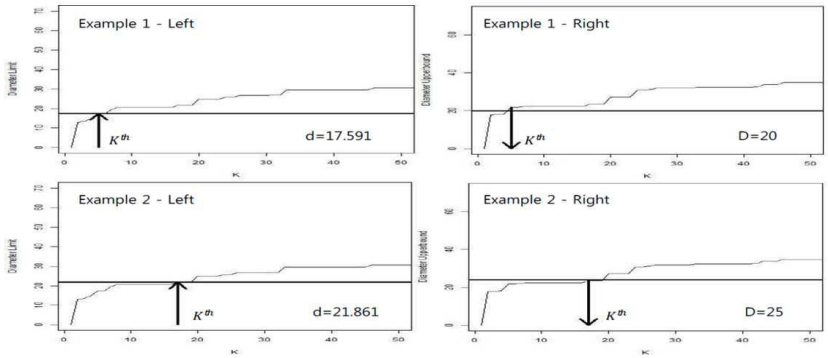
도면6



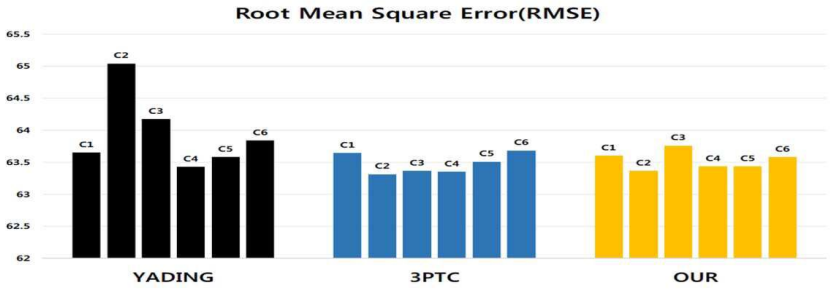
도면7



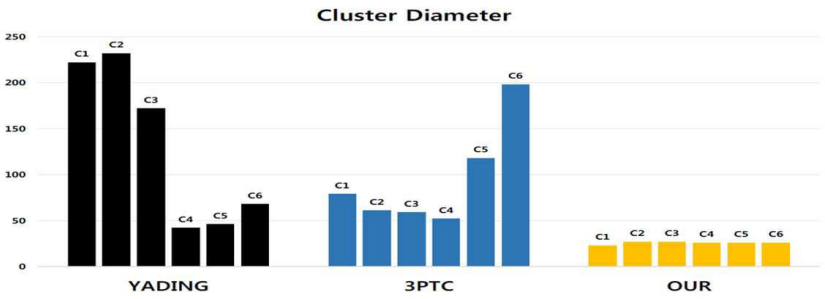
도면8



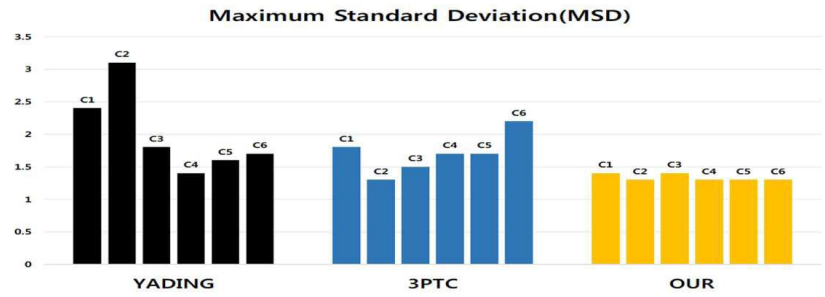
도면9



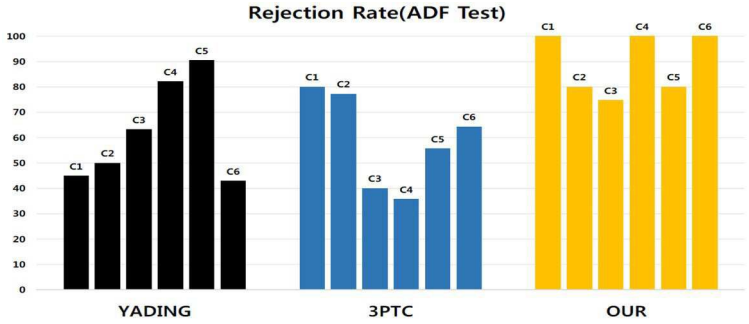
도면10



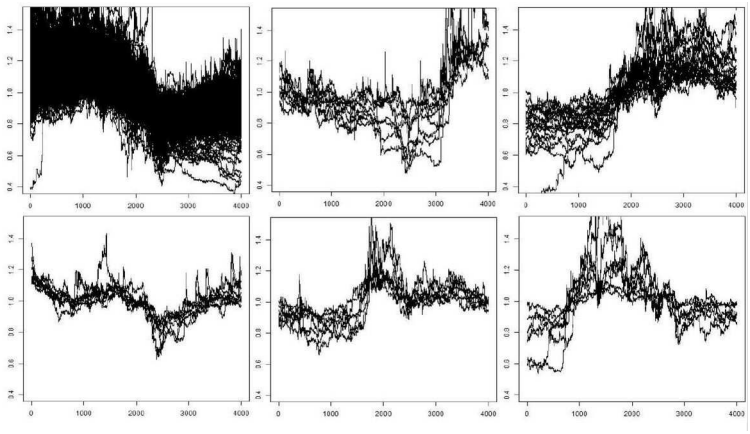
도면11



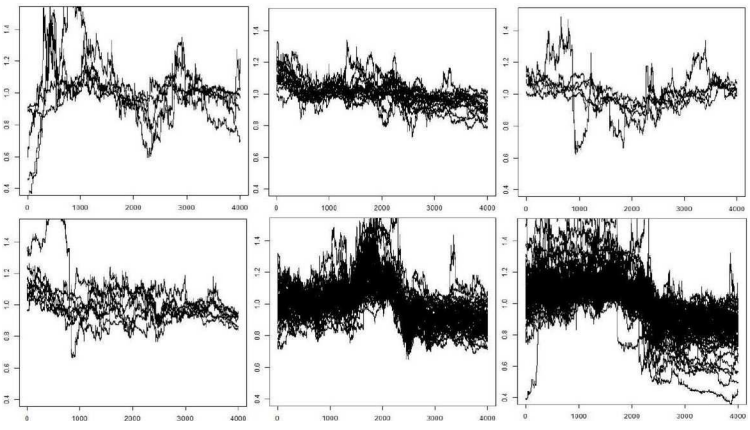
도면12



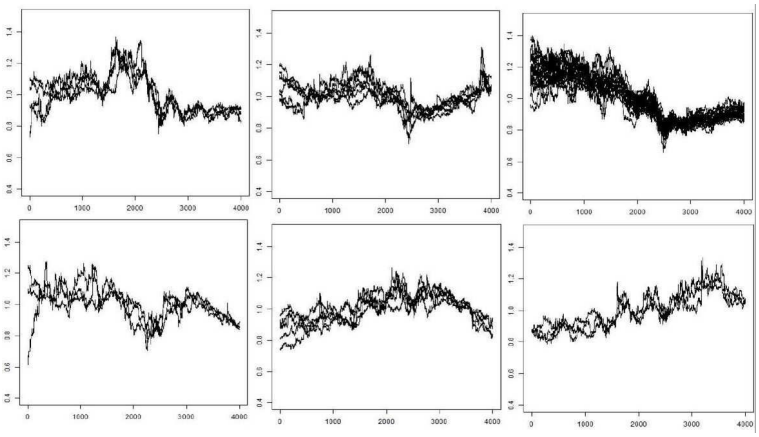
도면13



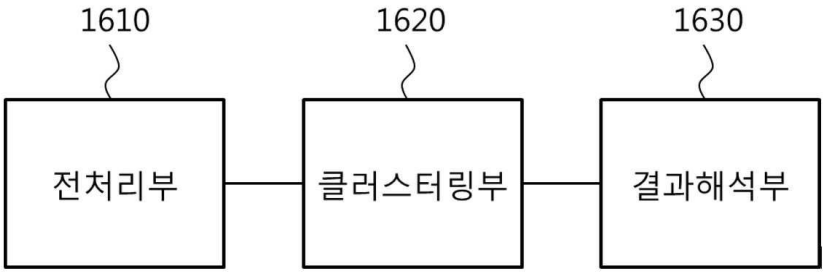
도면14



도면15



도면16



도면17